DOCUMENT RESUME

ED 477 927 TM 035 023

AUTHOR	Williamson, David M.; Bauer, Malcom; Steinberg, Linda S.; Mislevy, Robert J.; Behrens, John T.
TITLE	Creating a Complex Measurement Model Using Evidence Centered Design.
PUB DATE	2003-04-00
NOTE	39p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*College Students; Computer Assisted Testing; Computer Networks; *Computer Simulation; *High School Students; High Schools; Higher Education; *Measurement Techniques; *Task Analysis; *Test Construction
IDENTIFIERS	*Evidence

ABSTRACT

In computer-based simulations meant to support learning, students must bring a wide range of relevant knowledge, skills, and abilities to bear jointly as they solve meaningful problems in a learning domain. To function efficiently as an assessment, a simulation system must also be able to evoke and interpret observable evidence about targeted knowledge in a manner that is principled, defensible, and suited to the purpose at hand (e.g., licensure, achievement testing, coached practice). This paper concerns the grounding for the design of a simulation-based assessment of design and troubleshooting in the domain of computer networking. The application is a prototype for assessing these skills as part of an instructional program, as interim practice tests, and as chapter or end-of-course assessments. An evidence-centered assessment design (ECD) framework was used to guide the work. A total of 24 students in a computer networking class (high school and community college students) participated. An important part of this work was a cognitive task analysis designed to: (1) tap the knowledge computer network specialists and students use when they design and troubleshoot networks; and (2) elicit behaviors that manifest this knowledge. After summarizing the results, the paper discusses implications of this analysis, as well as information gathered from other methods of domain analysis, for designing psychometric models, automated scoring algorithms, and task frameworks, and for the capabilities required for the delivery of this example of a complex simulation-based assessment. (Contains 2 figures, 2 tables, and 27 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

The following paper was prepared for presentation at the April 2003 AERA/NCME program and is the copyrighted material of Educational Testing Service.

Running Head: COMPLEX MEASUREMENT MODEL

Creating A Complex Measurement Model Using Evidence Centered Design

ED 477 927

David M. Williamson

Malcolm Bauer

Educational Testing Service

Linda S. Steinberg

University of Pennsylvania

Robert J. Mislevy

University of Maryland

John T. Behrens

Cisco Systems

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D. M. Williamson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM035023

Correspondence should be addressed to:

David M. Williamson, Educational Testing Service, Mail Stop R-11, Princeton, NJ 08541

Unpublished work © 2003 by Educational Testing Service. Any opinions expressed herein are those of the authors and not necessarily of Educational Testing Service.

BEST COPY AVAILABLE

Abstract

In computer-based simulations meant to support learning, students must bring a wide range of relevant knowledge, skills, and abilities to bear jointly as they solve meaningful problems in a learning domain. To function effectively as an assessment, a simulation system must additionally be able to evoke and interpret observable evidence about targeted knowledge in a manner that is principled, defensible, and suited to the purpose at hand (e.g., licensure, achievement testing, coached practice). This paper concerns the grounding for the design of a simulation-based assessment of design and troubleshooting in the domain of computer networking. The application is a prototype for assessing these skills as part of an instructional program, as interim practice tests and as chapter or end-of-course assessments. An evidence-centered assessment design (ECD) framework was used to guide the work. An important part of this work is a cognitive task analysis designed to (a) tap the knowledge computer network specialists and students use when they design and troubleshoot networks, and (b) elicit behaviors that manifest this knowledge. After summarizing its results, we discuss implications of this analysis, as well as information gathered through other methods of domain analysis, for designing psychometric models, automated scoring algorithms, and task frameworks, and for the capabilities required for the delivery of this example of a complex simulation-based assessment.

Key words: Assessment design, cognitive task analysis, expertise, psychometrics, automated scoring.

Grounding A Complex Measurement Model Using Evidence Centered Design

INTRODUCTION

Decades of research and applied experience have honed the abilities of educational measurement practitioners to develop and implement a number of common, and relatively simple, assessment designs. Yet, even for these well-known designs, with firm theoretical grounding and substantial history, the conceptualization, development, and implementation of an assessment that meets professional standards continues to require substantial expertise, effort, and attention to detail. Cognitive science research has evidenced the complex nature of cognition when responding to even relatively simple items (e.g., Carroll, 1976; Klein et al., 1981; Whitely [Embretson], 1976). It follows that even in domains of interest well-served by relatively simple measurement models that the design effort, and the consequences of poor design, can be substantial.

While simple assessment designs suffice for many domains of interest, there are also many domains for which the nature of inferences to be made from an assessment demand complex task data to support these inferences. Assessment design for such domains demands greater complexity and innovation, and assumes risks inherent in exploring innovative approaches to measurement. These are risks that professionals may understandably be reluctant to accept when designing high-stakes assessments. Yet, the technological advances of the past decade have brought an unprecedented capacity for collecting complex and detailed assessment data in large scale administrations. This capacity is a potentially powerful tool for assessment, if well harnessed to the needs of measurement. Technology now exists for assessments to incorporate sophisticated and realistic simulations, which require examinees to draw upon a wide range of relevant knowledge, skills, and abilities as they solve meaningful tasks from the domain. A good simulation system, however, is not the same as a good assessment system (Melnick, 1996). To be effective as an assessment, a simulation system must be able to evoke, record and interpret observable evidence about targeted knowledge in a way that is principled, defensible, and suited to the purpose at hand (e.g., licensure, achievement testing, coached practice). It is not an effective strategy to design a simulation system and interesting tasks, and only then ask "How do you score it?" The foundation for sound assessment should be laid at the beginning and the principles of assessment design should guide every decision throughout the development process—tasks, scoring, psychometrics, simulators—so that the many elements perform in harmony to best serve the assessment purpose.

The field has been primed for innovative and complex assessments by the convergence of lines of research: cognitively-based task design; improved understanding of complex knowledge, skills and abilities (KSAs); and

innovative technology for delivery and managing assessments. Until recently, however, the grounding in sound design principles and corresponding methods for developing, administering, scoring, and maintaining such innovative assessments was not sufficiently developed to assure that resultant assessments would meet professional standards. Evidence Centered Assessment Design (ECD; Mislevy, Steinberg, & Almond, 2002; in press) leverages knowledge of cognition in the domain and sound design principles to provide a robust framework for designing assessments, be they simple and familiar or complex and innovative. The application of ECD to simple designs provides a rigorous and re-usable assessment blueprint, which facilitates developing assessments that meet professional measurement standards. For complex and innovative measurement needs, the ECD framework provides for assessment design that maintains an evidentiary focus, to guide the professional through the complexities of innovative design. This framework permits professionals to devote their efforts to the central design issues while avoiding unnecessary complexity and extraneous content (e.g. complete real-world fidelity, glitzy technology, etc.). The current work presents an application of ECD to assessment design for computer networking proficiency, designed for Cisco Learning Institute's (CLI) Cisco Networking Academy Program (CNAP), incorporating complex computer simulations. This paper presents a brief overview of ECD, describes the processes used to design the assessment (including a cognitive task analysis, or CTA), and discusses the implications of findings for design and implementation decisions.

ASSESSMENT CONTEXT

The CLI collaborates with high schools, community colleges, and vocational schools to provide education on the fundamentals of computer networking. The CNAP is four-semester curriculum teaching the principles and practice of designing, implementing, and maintaining computer networks capable of supporting local, national, and global organizations. Instruction is provided in classrooms as well as through on-line curriculum and activities and assessments are likewise conducted through classroom exercises and on-line testing. The CNAP uses the World Wide Web for both instruction and assessment administration and data maintenance. World Wide Web usage facilitates global access to educational resources, with approximately 150,000 students in 60 countries participating in the CNAP and an average of 10,000 administrations of online assessments per day. This high-volume global access presents both opportunities and challenges for educational research. Computer networking demands considerable technical knowledge as well as strategic and procedural expertise to become accomplished at common networking tasks. Despite the importance of these domain abilities current web-administered assessments consist of

multiple-choice items primarily assessing declarative knowledge. This fact prompted the lead CNAP curriculum designer, Dennis Frezzo (extracted with permission from *Assessment Pains*, an essay from a Cisco Learning Academy internal communication, May 2, 2000), to caution instructors not to depend on these assessments alone to evaluate students:

> *"[The current] online assessments ... are limited checks for understanding that will help the students get ready for that [certification] exam. But to produce students who can make real networks run, your assessment must be MUCH broader and deeper than any online assessment. ... Recall a primary goal of the program -- designing, installing, and maintaining networks. Quite frankly, the Assessment Server tests and [certification] test do not adequately test the complex problem-solving and manual set of skills required to maintain actual school networks. That is why the Instructor's Guide and Training model emphasize project-based, hands-on, lab-based, troubleshooting, "authentic", journal-and-portfolio-based assessments -- making cables, configuring routers and switches, building networks, wiring schools, all graded by rubrics."*

These comments reflect the widely held belief that the then-current on-line assessments were inadequate for determining student understanding of some of the most important aspects of networking ability. The absence of standardized assessment of critical elements of ability force a reliance on local evaluation efforts, which may be prone to substantial variability in curriculum emphasis and evaluation standards.

The limited scope of standardized assessment and the potential for substantial variability in student capabilities in critical areas of computer networking ability is being addressed by CLI, in partnership with ETS, through a redesign of the CNAP networking assessment program. This paper describes the initiation of a continuing collaboration among CLI, ETS, and the University of Maryland to redesign the networking assessment program to more appropriately measure critical computer networking abilities. This redesign leverages current simulation technology and remote connection capabilities to produce an online assessment exercising the *cognitive* aspects of network design, implementation, and troubleshooting. The initial outcome is a prototype assessment, called NetPass, using network simulations with realistic interactive tasks to measure students' abilities and provide targeted educational feedback. This feedback includes reporting on the students' knowledge of networking, their mastery of various networking skills, their ability to carry out procedures and strategies for networking tasks, and their misconceptions about network functionality and operational procedures. An ECD process was employed in this

redesign, providing the framework necessary to meet the design needs for such a complex computerized assessment in a technical domain.

EVIDENCE CENTERED ASSESSMENT DESIGN

The inherent complexity of the computer networking domain demands meticulous care in assessment, particularly in the design of simulations to evoke behavior that provides targeted evidence about key skills and knowledge. Furthermore, the design must provide for principled interpretations of elicited behavior and how it constitutes evidence that suit the purpose of the assessment. Psychometric models for such complex assessments are feasible only after substantial integrative design work prior to the implementation of the statistical model. The ECD framework ensures that appropriate evidence is gathered for assessment inferences, thus supporting the validity of those inferences by requiring explicit connections between the claims to be made about students, the evidence supporting these claims, and tasks that elicit such evidence from students. Evidence Centered Design (Mislevy, Steinberg, & Almond, 2002, in press) emphasizes the explicit definition and full explication of design elements, including their relation to assessment purpose and other design components, in each step of the design process. As such, ECD is a systematic approach to design that requires consideration and explicit definition of measurement constructs and evidence that would support the types of inferences to be made from the assessment (Mislevy, 1994). As is true with most systematic design methodologies, ECD requires greater initial effort, but yields greater ability to understand the constructs being measured and the impact that changes in assessment requirements and designs will have on subsequent design and development stages (see Steinberg et. al., 2000) as well as producing a re-usable blueprint for assessment design.

By virtue of emphasizing systematic consideration of targeted inferences, the nature of evidence required for such inferences, and the explicit relationship between design components and the delivery processes, the final assessment design from ECD also provides specifications for implementation and delivery. The Conceptual Assessment Framework (CAF) for ECD is depicted in Figure 1. It consists of four fundamental models: the student model, representing student KSA structure and forming the foundation for claims that will be made about the student on the basis of assessment evidence; task models, specifying content and construction of tasks in order to provide the necessary evidence to support claims about the student; evidence models, specifying how evidence from assessment tasks inform student model variables in support of specific claims about the student; and an assembly model,

specifying the strategy used to select and present tasks to a student (Almond, Steinberg, & Mislevy, 2001). The CAF addresses a sequence of questions posed by Sam Messick (1994) that represent the foundation of assessment design:

- "What complex of knowledge, skills, or other attribute should be assessed?"
- "What behaviors or performances should reveal those constructs?"
- "What tasks or situations should elicit those behaviors?"

[[Insert Figure 1 about here]]

"What complex of knowledge, skills, or other attribute should be assessed?" The student model is comprised of variables representing characteristics of the examinees—knowledge, skills and abilities—that are the focus of the inference process and that determine the outcomes (e.g. licensure, placement, diagnostic feedback, or some combination) consistent with assessment purpose. The student model represents knowledge about an examinee's possession of these KSAs based on assessment evidence. The extent and functionality of a student model can vary widely by assessment purpose, and include multiple student models to support multiple purposes from a common assessment (e.g. one to provide for outcomes such as an overall score and another to provide for educationally relevant feedback).

"What behaviors or performances should reveal those constructs?" An evidence model expresses how the observable features of examinee work products from a task constitute evidence about student-model variables. The evidence model is made up of two components: (a) the evaluation component, which consists of the *rules of evidence* describing the evidence extraction process, in which features of the work product are identified as observable variables that constitute evidence for inferences about students, and (b) the *statistical, or measurement, model*, which specifies how information provided by these observables should influence belief about the values of student model variables.

"What tasks or situations should elicit those behaviors?" Task models describe characteristics of the assessment tasks intended to elicit particular types of examinee performance. A task model provides a framework for characterizing and constructing situations with which a candidate interacts to provide evidence about targeted aspects of knowledge. The task model specifies key elements of 1) performance situations, 2) material presented to the student, and 3) student work produced in responding to the task (see Bennett & Bejar, 1998 for a discussion of the importance of this and other considerations). A typical assessment uses not one but many task models, each capable of generating many tasks according to model specifications. Tasks generated from such models share similarities in task requirements and the student work produced in response. Both task models and instantiations of

items from them play a vital role in automatic item generation efforts (Bejar, 2002; Irvine & Kyllonen, 2002; Mislevy et al., 2002, in press; Mislevy et al., 1999).

In an implemented assessment, the student model(s) accumulate and represent current beliefs about the targeted aspects of examinee proficiency that are expressed as student-model variables. The evidence models identify the key features of examinee behaviors and work products that provide evidence about aspects of proficiency in the student model (Steinberg & Gitomer, 1996), and use a psychometric model to express how the evidence relates to student-model variables (Mislevy, 1994). The task models guide the design of situations engineered to evoke responses that provide appropriate evidence for the claims relevant to the purpose of the assessment (Almond, Steinberg, & Mislevy, 2001). An example of the use of these models to develop a design rationale for simulation-based problem-solving in the domain of dental hygiene is provided by Mislevy, Steinberg, Breyer, Almond, and Johnson (1999; in press).

UNDERSTANDING THE DOMAIN

Under ECD, the design process begins by specifying the purpose of the assessment. The design team uses this understanding as the origin for analyzing the domain, with the intent of identifying and specifying domain-relevant *claims* to be made about students on the basis of the assessment. These explicit purpose-related claims guide development of the assessment to ensure that the observable evidence provided by the assessment tasks is both relevant and sufficient to support such claims. In particular, the claims provide the foundation for the identification and specification of proficiencies students must possess, and that the assessment must elicit evidence of, in order to make valid and relevant inferences about the student. This catalog of domain proficiencies focuses the domain investigation toward a full understanding of the nature and extent of these critical proficiencies. The explicit relationship between the claims to be made, the proficiencies that students must possess to meet these claims, the nature of data that provides sufficient evidence of these proficiencies, and the critical features of situations that can elicit such evidence, provides the evidential framework for the development of an assessment capable of producing valid and relevant results.

The primary purpose of the NetPass prototype assessment is to provide a means for CNAP students to obtain a standardized evaluation of their networking skills, to complement current multiple-choice tests assessing declarative networking knowledge. This purpose implies assessment of strategy, efficiency and other aspects of operational computer networking performance that are difficult to capture with multiple-choice questions. The goal

is to both assess networking proficiency and to provide educational benefits to the student through targeted diagnostic feedback from their performance. A primary goal of the assessment is to provide access to evaluation on a single CNAP standard for networking ability regardless of physical location of the student as well as instructionally relevant diagnostic feedback from highly monitored performance. This service is expected to be a substantial improvement over reliance on potentially varied standards provided by regional instruction alone. As part of the specification of assessment purpose it was decided to limit the scope of NetPass to the most critical aspects of computer networking—network design, network implementation, and network troubleshooting—and to therefore exclude routine planning and maintenance from the prototype assessment.

This first stage of assessment design fulfills several purposes: defining the scope and intent of the assessment, documenting and delineating the specific claim-based utility of assessment results, and specifying the rationale and assumptions underlying these decisions. While these design components are repeatedly revisited and revised during the design and development process as a result of further investigations, these decisions set the stage for documenting the claims to be made on the basis of assessment results.

Delineation of the assessment purpose allowed the design team to begin developing claims to be made about students on the basis of the assessment. Specific claims were developed through domain research about the knowledge, skills and abilities necessary for computer networking performance at the ability levels targeted (on the basis of the previous delineation of assessment purpose) for the assessment. This process required multiple sources of information, including CNAP curricular material, subject matter experts (practitioners and instructors), existing assessments, computer networking documentation, educational texts and courseware, and prior research in related fields. As the four-semester CNAP curriculum was developed from a similar process focusing on success as a computer networking practitioner it was not surprising to find that the resultant claims overlapped substantially with the CNAP curriculum.

Claims are specific statements about the examinee's knowledge, skills, or abilities to be made on the basis of observed evidence from the assessment. The full list of domain-based claims is developed through an iterative process of claim addition, culling, and hierarchical reorganization. The hierarchical structure can reflect choices in degree of specificity with which claims are established and ultimately, can be reflected in the hierarchy of the student model for the assessment. For example, three claims that emerged from the NetPass analysis were:

1. Students can use a systematic approach to identify and solve network problems

2. Students can identify the cause of connectivity problems at the physical, data link, and network layers of the OSI model
3. Students can use TCP/IP utilities to troubleshoot network connectivity problems.

The first claim is a general statement about troubleshooting ability while the second claim is a subclaim specifying a component ability needed for the first claim, in this case the ability to troubleshoot a common network symptom, namely, loss of connectivity. The third claim is in turn a subclaim of the second claim and is one of the subskills needed to have the ability described in the second claim. The full set of claims in the NetPass design is a hierarchical tree with four tiers, starting with a high level statement about students' networking abilities and branching down to more specific abilities and subskills. Ultimately, these hierarchical claims are partially represented in the final form of the student model for the assessment.

The claim hierarchy provided explicit targets guiding the design process delineating a sufficient body of supporting evidence. While the full catalog of claims is extensive, this paper presents a limited portion of claims focusing on knowledge, skills and abilities consistent with the third semester of CNA curriculum for discussion purposes. Claim development and hierarchy was facilitated by both documentation resources and subject matter expert (SME) identification of undocumented skills and knowledge implicit in expectations about course laboratory performance. After several iterations of examination and refinement by SMEs the final list of claims formed the basis for determination of the nature and extent of evidence required to support these claims about students.

Once the claims are established the next design task targets the specification of evidence that must be observed in examinee performances to provide sufficient support for these claims about examinees. The documented relationship between the evidence observed from examinee work and the claims to be made about the examinee provide a substantial degree of support for the validity and utility of the assessment for its intended purpose. As assessment design continues both the claims and the evidential requirements are revisited and revised to reflect the dynamic elements of the design process (e.g. subsequent task development, etc.) and to ensure that the evidence being provided by the assessment is sufficient to support each claim.

On the basis of established claims, the design team worked with SMEs to develop descriptions of ways knowledge is represented in the domain and the features of these representations that can provide evidence about the knowledge and skills possessed. An example of the types of representations corresponding to Claim 2 (discussed above) and a table expressing how features of those representations provide evidence for Claim 2 is provided as Table 1. In Table 1 a *network diagram* is a common way for networking professionals and students to describe the

structure of a network and it displays network devices, connections between network devices, and network address information. A *log file* consists of the sequence of network commands used to configure or troubleshoot a network. A *configuration file* describes the state of variables associated with a network device (typically a router, for which IP addresses and other characteristics may be specified).

[[Insert Table 1 about here]]

The design team, again with SMEs, identified the features of work products that provide evidence regarding Claim 2, which is provided as Table 2. Scoring features, identifiable characteristics of an examinee solution that can be used as evidence to support claims, are provided in the first column. The second column provides the general category containing the feature and the third column provides the link to representations in Table 1 that are capable of representing the feature. For example, the second feature, "identification of network problems", provides critical evidence for the claim because identifying the specific network problem (a bad network address, wrong communication protocol, etc.) is essentially identifying the cause of the loss of network connectivity. This evidence would be contained in multiple representations: the log file; configuration file; and worksheet. The log file would record the commands the student entered while conducting troubleshooting to investigate and fix the connectivity problem. Similarly, a configuration file would provide the subsequent configuration of the network after the student's interventions during troubleshooting. Finally, a worksheet could record faults identified by the student. As such, Tables 1 and 2 describe the evidence that can be gleaned from different representations in support of specific claims.

[[Insert Table 2 about here]]

Explicating the relationship between claims and different types of evidence contained in various representations also involved specifying the necessary characteristics of assessment task situations. Documented task characteristics describe the necessary task elements that enable a student to employ targeted knowledge and skill, and, given these situations, the ways he or she could act to display that knowledge. The components of the computer network, their associated representational systems, and the beginning and ending network states were particularly important in the specification of task performance situations.

Consideration of the necessary characteristics of assessment tasks includes addressing the method of administering such tasks, since this can impact the nature of evidence collected. Particular administration requirements were thus derived from the nature of claims and evidence for NetPass and the goal of assessing student ability to interact effectively with a computer network while addressing complex networking tasks. NetPass required

networking tasks that, if not performed on an actual computer network, must have an interactive computer interface which functions with a high degree of fidelity to an interface for a real computer network. Additionally, on-line web-based delivery was chosen as the administration vehicle in order to provide a universal assessment standard in a practical and readily accessible manner for remote learning locations.

Design decisions such as these have the effect of both targeting the primary goals of the assessment and simultaneously precluding other potential design features. For example, by meeting the purpose of assessing computer network interaction skills through widely accessible internet administration, the assessment must exclude a direct assessment of the student's ability to examine the physical layer of the network (e.g. to check that network cables are connected during troubleshooting activities, to physically install network cards in computers, etc.). As proficiency in physical layer operations was not a high priority goal of the assessment, this was an acceptable constraint on the assessment design. Such tradeoffs illustrate how the relationship among claims, the evidence that support those claims, and the characteristics of tasks and administration characteristics must be repeatedly revisited and reexamined during the design process. This work was a precursor for the cognitive task analysis, which provided the level of detail needed to fully define the remaining elements of the assessment design.

Cognitive Task Analysis

Critical to the NetPass assessment design was a cognitive task analysis (CTA). A CTA is a disciplined process of investigating the knowledge structures and strategies that individuals at targeted levels of ability use to solve specific types of tasks, and observable evidence of those structures and strategies (e.g. Steinberg & Gitomer, 1993). While a job analysis is typically concerned with the frequency and importance of domain tasks performance, a CTA focuses on identifying knowledge and strategies people use to address those tasks. A CTA seeks to expose (a) essential features of task situations for eliciting certain behaviors; (b) internal representations of task situations; (c) the relationship between problem-solving behavior and internal representation; (d) processes used to solve problems; and (e) task characteristics that impact problem-solving processes and task difficulty (Newell & Simon, 1972).

With the ultimate objective of designing optimal assessment tasks, CTA methods were adapted from expertise literature (Ericsson & Smith, 1991) to capture and to analyze the performance of CNA students at different known levels of expertise, under standard conditions, across a range of tasks. CTA conducted in the service of assessment design will focus on skills, behaviors, knowledge structures and cognitive strategies that are directly

related to the assessment purpose, rather than on the general understanding of cognition in a domain that characterizes CTA in cognitive science. The preceding steps in domain analysis for computer networking establish a fundamental understanding of the domain and the intent of the assessment, thereby directing the focus of the CTA on specific areas of task performance most critical to the assessment claims, and thus improving the efficiency of the CTA in support of the assessment purpose.

The CTA for this project was designed to (a) tap the knowledge and strategies used by CNAP students of various ability when designing, implementing, and troubleshooting networks, and (b) identify observable behaviors that manifest this knowledge and strategy at various levels of proficiency. The CTA was designed to flesh out the assessment design through further identification of knowledge and skills required for network design, implementation, and troubleshooting, with respect to the assessment purpose of low-stakes (learning) assessment with supplementary feedback. The assessment claims most relevant to the third semester of CLI curriculum, the primary target of NetPass, represented the focus of the CTA. These claims cover three skill areas: network troubleshooting, network configuration, and virtual local area network (VLAN) design.

Materials

CTA requires stimulus tasks for students to work through during think-aloud protocols. The tasks developed for the CTA were constructed to allow subjects to demonstrate their understanding and ability in specific areas of computer networking, primarily design, implementation, and troubleshooting. To ensure that the full range of ability could be demonstrated in the CTA, three tasks were developed for each area of emphasis (design, implementation, and troubleshooting) with each task targeting a different degree of knowledge and skill (high, moderate, and low), resulting in a total of nine scenarios. These degrees (high, moderate, and low) of targeted knowledge and ability were identified with reference to the domain content and tasks typically associated with the 3rd semester of CNAP curriculum. Therefore, it would be expected that for the NetPass prototype target population of CNAP students who had successfully completed the 3rd semester there would be one easy task, one moderately challenging task, and one difficult task available for each of the three areas of emphasis (design, implementation and troubleshooting). For each task, actual Cisco networking equipment was set to specified initial conditions. Students then solved the tasks as described in general below for the three areas of emphasis.

Network troubleshooting. The student is introduced to an existing network, with specified properties and meant to perform in predetermined ways. User reports of certain failures are provided. It is the student's task to determine the fault(s) and fix them.

Network configuration. The student is presented with the design specification of a network. The student configures a provided pod of routers to reflect the design specification.

VLAN design. The student is presented with user requirements and constraints for designing a local or wide area network involving a VLAN. The student develops a design to satisfy these requirements and constraints.

Network Diagram. In association with each of the above three activities, the student produces a diagram of the structure of the network, the relationships and connections between network elements, and the network functionality for the network they designed, implemented, or troubleshot. This activity was required for each scenario based on the analysis of the domain, which indicated that the understanding of physical systems and their relationships in a computer network is a critical component of networking ability.

Participants

A total of 24 students in the 3rd semester of the CNAP curriculum were recruited from multiple academy locations in community colleges and high schools in North Carolina, Georgia, and Montana to serve as CTA participants. These students were selected by their instructors to represent three levels of ability among those completing 3rd semester CNAP curriculum (8 lower, 8 average, 8 high)[1]. The general ability evaluations of the instructors were corroborated through the use of a multiple-choice pretest measure prior to participation in the CTA.

Method

Each participant took a pretest and solved four CTA scenarios; one pair from each of two of the three content areas. Participants were asked to think-aloud as they solved the scenarios and their processes were recorded. After participants completed the set of four scenarios, they were asked to recall problem-solving rationales for each scenario using a structured retrospective protocol in which a researcher reviewed their solutions with them. The ordering of scenarios were assigned to control for sequential learning effects and difficulty order effects to the extent

[1] Note that participant ability categories differ from CTA analysis in cognitive science, which typically examines differences between novices and acknowledged domain experts. This difference in choice of participant ability

possible. Participants were assigned to tasks consistent with instructor estimates of student ability based on the assumption that such targeting of task difficulty to ability would yield more useful results than random or stratified assignment. As a result of the previous phases of domain analysis, the design and troubleshooting areas were believed to be more critical to the assessment design. Moreover, they offered a potential for more variation in students' performance than network implementation, which tends to rely on a small set of fairly standard procedures. Therefore, more participants were assigned to troubleshooting and design scenarios than implementation scenarios: seven pairs of observations for troubleshooting and seven pairs for design for each participant ability level, and two pairs of observations for implementation for each participant ability level.

Analysis

The CTA data consisted of transcripts of talk-aloud solutions, log files of all computer workstation commands during task completion, and diagrams and calculations produced as participants solved the tasks. This data was analyzed and discussed by a team of ten researchers, computer networking instructors, and subject matter experts with the intent of identifying recurring patterns that distinguished performance scenarios and ability levels. In this process prior design work was again called upon to suggest hypotheses for investigation in talk-aloud protocols and resultant log files and diagrams. These identified patterns served as the foundation for defining re-usable observed variables from assessment tasks for use in the evidence models.

CTA Results

Troubleshooting Tasks

Examination and discussion of student troubleshooting protocols and associated log files ultimately resulted in the classification of various commands and sequences of commands into four major categories:

1) *Actions associated with gathering information about router configuration.* These commands provide information about the state of a router and network connectivity (e.g. the "show" command and its variants).

---

range is in deference to the assessment purpose defined in prior stages of design, illustrating how prior design work influences subsequent design decisions.

2) *Actions associated with changes to router configuration to fix faults*. These commands institute changes to a router (e.g. adding a clock signal, setting router protocol, setting IP addresses on interfaces, etc.).

3) *Actions associated with testing network behavior after fixes*. These commands also provide information about the state of the network but for post-fix testing these commands are typically directed at network connectivity (e.g. "ping" and "telnet") rather than provision of router configuration information (e.g. "show"). However, some of these commands can overlap with those described for information gathering above. Therefore, in establishing whether actions are information gathering or post-fix testing the prior fault-fixing actions must also be referenced.

4) *Actions associated with getting information about commands*. These commands do not address the network but instead access available information about commands or command targets (e.g. uses of the help system, "?" command).

Through this command classification structure it was possible to analyze troubleshooting activities with respect to the frequency, character, and appropriateness of actions based on these command classifications.

These characterizations and classification of commands facilitated subsequent SME examination of protocols and command patterns, ultimately resulting in identification of three patterns of solutions. It must be emphasized that particular students did not necessarily exhibit consistent patterns on all tasks, and that the patterns were correlated with, but not identical to, the instructor designated ability levels. That is, the following are descriptions of *behaviors*, not of *students*. The characteristics that differentiated the three performance patterns could be summarized by means of two main categories, which themselves are composed of a hierarchy of contributing components according to the following structure:

a) Correctness of Procedure
   i) Procedural Sequence Logic
      (1) Sequence of targets
      (2) Sequence of actions
   ii) Efficiency of Procedure
      (1) Help usage
      (2) IOS syntax

(3) Volume of actions

b) Correctness of Outcome

i) Error Identification

ii) Error Over-Identification

The typical patterns in troubleshooting solutions were the following:

- **Troubleshooting Pattern A.** The student followed unsystematic and inefficient troubleshooting procedures, rarely found and fixed faults, and used the help system extensively to guide command use.
- **Troubleshooting Pattern B.** The student found and fixed faults correctly and efficiently.
- **Troubleshooting Pattern C.** The Student followed standardized troubleshooting procedures rather than a sequence of procedures tailored to the network in question. These patterns of solutions fell between Patterns A and B in their overall success on the troubleshooting tasks. These patterns follow a more prescribed set of procedures and usually take more steps to isolate each problem than a Pattern B solution. The student attempted to be systematic in their testing of the network, but sometimes flailed to identify the fault(s) in the network, or did so through a long and circuitous series of actions. They correctly fixed some of the network faults, but they also fixed faults that were not actually present.

Troubleshooting pattern A: correctness of procedure.

These solutions generally lack direction in their troubleshooting; that is, the solutions were undertaken with a poor procedural sequence logic and with poor efficiency of procedure. Poor procedural sequence logic was evidenced in these solutions by student exhibition of a random and haphazard investigation of network areas (the 'targets' of investigation) without leveraging clues regarding suspect areas of error location from the scenario statement or following any accepted standard sequence of investigation of network components. In addition, poor procedural logic was evidenced by student failure to follow a reasonable sequence of actions when investigating targets, typically by performing such actions as changing the state of the network without assessing the need for the change, or by failing to assess whether network function was successfully altered after making a change to the network.

Efficiency of procedure was poor in these solutions, with characteristic patterns of actions displaying excessive help usage, poor mastery of IOS syntax, and a high volume of actions as they worked through the scenario.

With regard to help usage these solutions evidenced repeated commands requesting system help, showing a high reliance on the help features to guide them through the scenario. Such solutions also tended to have a substantial number of invalid commands entered resulting in a poor demonstration of understanding the IOS syntax needed to perform network operations. Furthermore, these solutions were characterized by a high volume of actions, including repeated "show running-configuration" commands or other information gathering commands, especially looping through different routers repeatedly asking for information without actually fixing anything; and gathering information on aspects of the network that don't need to be examined in light of the information students have obtained thus far and a reasonable understanding of the curriculum. For example, the problem statement of the difficult troubleshooting scenario makes it clear that the fault lies with an access control list, but some students look elsewhere, for example, with repeated show interface commands. These solutions are characterized by inefficiency in each of these areas when working through the troubleshooting scenario.

Troubleshooting pattern A: correctness of outcome.

These solutions generally lack a satisfactory outcome of the student troubleshooting efforts by both failing to remedy malfunctioning aspects of the network and by 'fixing' aspects of the network that are not problematic (e.g. make unnecessary changes to the configuration of the router, apply a clocking signal to an interface even though one was present in the original configuration, etc.). As a result of this tendency to fail to identify and implement solutions for network problems and the tendency to implement changes to functional areas of the network their correctness of outcome was poor for these solutions.

Troubleshooting pattern B: correctness of procedure.

These solutions demonstrated very direct troubleshooting procedures that follow an appropriate procedural sequence logic with high efficiency of procedure. High degrees of procedural sequence logic were evidenced in these solutions through a direct and targeted investigation of malfunctioning network components without excessive investigation of functional elements. These network components were investigated with an appropriate sequence of actions by following a logical pattern of investigation, implementation of network changes, and testing of the network for the effectiveness of the change. Furthermore, a high efficiency of procedure was characteristic of these solutions as the help features were seldom used, syntax errors in command entry were uncommon, and the scenario was addressed with a low volume of actions.

Troubleshooting pattern B: correctness of outcome.

Pattern B solutions demonstrated a tendency for students to address actual faults in the network, without making 'fixes' to areas of the network that were functioning properly, and to address the existing faults correctly, thus demonstrating a high level of performance on the correctness of outcome.

Troubleshooting pattern C: correctness of procedure.

These solutions generally demonstrated some direction in their troubleshooting, though typically floundering at different points in their troubleshooting process; that is, the solutions were undertaken with a moderate procedural sequence logic and with moderate efficiency of procedure. A moderate degree of procedural sequence logic was evidenced in these solutions by frequent student exhibition of a sub optimal strategy of targeting areas of the network for investigation called serial elimination; starting at one end of the network and working systematically through to the other end of the network rather than capitalizing on an understanding of the scenario information thus far to carry out a more efficient strategy, such as space splitting; in which whole sections of the network are eliminated from the possibility of containing the error. There were also occasional lapses into undirected sequences of actions, similar to Pattern A solutions, but for only portions of the problem. In addition, moderate procedural logic was evidenced by students following a general but inconsistent tendency for investigation of the network function prior to making changes and then following up the change with confirmation of the function, as well as some tendency to make changes without investigation or confirmation of the impact of the change. Finally, a moderate degree of efficiency of procedure was characteristic of these solutions as the help features were occasionally needed, syntax errors were somewhat prevalent, and the scenario was addressed with a substantial volume of actions.

Troubleshooting pattern C: correctness of outcome.

Pattern C solutions demonstrated a tendency for both addressing some exiting faults successfully and for a tendency to 'fix' faults in the network which were not actually present. The final result of which is that some, but not all, existing faults were remedied and some additional unnecessary fixes were implemented in the network, resulting in a final outcome at a moderate level of performance.

Design Tasks

Based on the examination of CTA results, several patterns of performance emerged which distinguish varying degrees of quality in Design solutions. The basic distinction among these design tasks was with the *Correctness of Outcome*, which focuses on whether the resulting network design is functional, correct, and complete.

In professional practice there are two aspects to Correctness of Outcome: the Functionality of Design, which is a measure of the extent to which the network serves its intended purpose, and the Efficiency of Design, which considers aspects which affect network performance, such as the number of components used, the cost of components, and the maintenance and performance implications of the selected components. Both aspects are important in professional practice, but since the vast majority of professional designs are functional they differ mainly in their efficiency. In contrast, efficiency is not a major factor discriminating among the student-generated designs in the relatively simple problems addressed in third-semester CNA curriculum; for the most part, they either satisfy the requirements or they don't. Therefore, only the Functionality of Design was found to be relevant in analyzing the student design solutions.

Functionality of Design can be decomposed into two parts: Core Requirements and Peripheral Requirements. The Core Requirements represent the critical elements of the function of the network design that must be in place for the network to meet even the rudimentary intent of the network function (for example, in a VLAN design a student must have indicated the right broadcast domains). The Peripheral Requirements are elements that are required for network performance, but which are not central to the purpose of the network or the intent of the task (for example, having an appropriate number of signal repeaters to account for distance). The characteristics distinguishing patterns of performance for Design tasks are:

a) Correctness of Outcome

i) Functionality of Design

(1) Core requirements

(2) Peripheral requirements

On this basis three patterns of performance were distinguished among the solutions to the design problems in the CTA:

- **Design Pattern A.** This pattern was characterized by designs that result in a network that do not meet the core operational requirements for functionality of design.
- **Design Pattern B.** This pattern was characterized by designs that result in a network that meets all the core operational requirements and many of the peripheral operational requirements for functionality of design.
- **Design Pattern C.** Designs that resulted in a network that meets all, or very nearly all, of the core operational requirements for functionality of design.

Design pattern A: correctness of outcome.

These solutions produced designs that do not meet the core operational requirements for functionality of design. Given the failure to meet the core operational requirements, the degree of success in designing for the peripheral elements of the network is moot.

Design pattern B: correctness of outcome.

The student created a design that possesses all, or very nearly all, of the core functionality of design called for in the problem statement. In addition, the student design also exhibits many of the peripheral functionality requirements either explicitly or implicitly requested in the problem.

Design pattern C: correctness of outcome.

These solutions met all, or very nearly all, of the core operational requirements for functionality of design. However, there was little or no representation of the important peripheral features and operation of the network.

Implementation Tasks

Review of the CTA results and discussion with SMEs established that the overall framework of observations appropriate for implementation is similar to that for troubleshooting. Both the way in which students perform the configuration (correctness of procedure), and the quality of the resulting configuration (correctness of outcome) are important high-level behavioral characteristics. The same types of actions (gathering information, making changes or adding to the configuration, testing, and getting help) that apply to troubleshooting activities apply to implementation as well. Compared to troubleshooting, however, the implementation tasks demand greater emphasis on the network configuration changes than on gathering information and testing the network function. The characteristics that differentiated the patterns could be summarized by means of these two main categories, which themselves are composed of contributing components according to the following structure:

a) Correctness of Procedure
   i) Procedural Sequence Logic
   ii) Efficiency of Procedure
      (1) Help usage
      (2) IOS syntax
      (3) Volume of actions
b) Correctness of Outcome

Correctness of Procedure has the same two components as for troubleshooting: efficiency of procedure and procedural sequence logic. Efficiency is largely the same as in troubleshooting, focusing on the overall number of actions, use of help, and correct syntax of IOS commands. There are some optimal patterns of behaviors relating to sequence of procedure, but they are less constrained than in troubleshooting; often, the order in which commands are used does not matter in configuration. It does matter in some cases, however, including (1) the order of commands in constructing an access control list, and (2) the sequencing of actions that are used to configure and test connectivity (e.g. it is necessary to set up the routing protocol and interfaces on the routers before testing for connectivity).

Correctness of outcome consists of getting the configuration right – however, some components of the configuration are more difficult than others (e.g. access control lists). In addition, there are some aspects of the configuration (e.g. setting passwords) that students should perform even if the scenario only explicitly mentioned higher-level requirements (e.g. security issues, in the case of passwords). The categories of behavior that reflect these observed differences in patterns of performance are:

- **Implementation Pattern A.** This pattern of performance was indicative of solutions in which students experienced substantial difficulties in configuring network devices.
- **Implementation Pattern B.** This pattern represented those solutions in which students implemented correct solutions in a straightforward manner.
- **Implementation Pattern C.** This pattern was characteristic of solutions in which the student experienced some difficulty in properly configuring network devices.

Implementation pattern A: correctness of procedure.

Solutions with this pattern of performance had inappropriate procedural sequence logic evidenced by tendencies to test the network configuration at inappropriate times or to forego any testing of the configuration whatsoever. These students also conducted tasks in inappropriate sequences, such as working on access control lists before verifying that their initial configuration was working properly.

Similarly to the pattern for troubleshooting, with regard to efficiency of procedure this pattern of performance exhibited extensive reliance on system help to find the correct commands to execute in implementing the network. In addition, these solutions contained many syntax errors, and far more information was gathered about the network than was actually needed to conduct the task, resulting in the use of an excessive number of commands to complete the task.

Implementation pattern A: correctness of outcome.

Solutions with pattern A had many "sins of omission" in which the student omitted important parts of the configuration (e.g. failing to include a network address command, leaving an interface unconfigured, omitting one of the addresses when enabling a router protocol). Protocols of this type would also configure network components incorrectly (e.g. using the wrong network address when enabling a router protocol). For this pattern these types of errors could occur on many aspects of the configuration, including:

- Naming the router
- Setting passwords
- Host tables
- Descriptors on interfaces, and router tables
- Defining protocols and static routes
- Enabling web browser capabilities

In addition, there were often problems with the access control lists, with this pattern often having the access control list applied to the wrong interface, the wrong wild card mask set, or with the statements in the wrong order.

Implementation pattern B: correctness of procedure.

In these solutions there was a demonstration of appropriate procedural sequence logic as the students tested the network configuration at appropriate times during the implementation process. In addition, they verified that their initial configuration was working properly before establishing access control lists. These solutions also evidenced a high degree of efficiency of procedure as there was rare use of help to find the correct commands, few syntax errors, and information about the network was gathered judiciously. Furthermore, the total number of actions taken to conduct the task was minimal with this pattern of solution.

Implementation pattern B: correctness of outcome.

Pattern B solutions were indicative of networks that were completely or almost completely correct in their implementation, including correct implementation of access control lists.

Implementation pattern C: correctness of procedure.

The procedural sequence logic of these solutions were characterized by a generally appropriate use of the "show" command to gather information about the network. Within this pattern the solutions did show appropriate tests of the network configuration as the configuration process continued. However, these network tests were often

sub optimal in that the student would conduct excessive testing or refrain from any testing until after making many changes to the configuration. In virtually all cases, however, these solutions demonstrated that the initial configuration was verified as working properly before work on access control lists began. These solutions also indicate some tendency to overlook certain configuration elements initially, requiring the student to eventually return to the router configuration mode to remedy portions previously omitted. Yet with this pattern of activity, as opposed to pattern A, the students did return to correct portions of the network configuration not previously addressed.

The efficiency of procedure for these solutions was likewise moderate, with the occasional use of help to find the correct commands, the presence of some syntax errors in the implementation process, and an overall tendency to use a number of actions to complete the work falling between those for Pattern A and Pattern B solutions.

Implementation pattern C: correctness of outcome.

Pattern C configurations were typically correct, for the most part, but would also typically contain some of the errors that comprise pattern A solutions. In particular, there may be some difficulties with the implementation of the access control lists in pattern C solutions, though not to the extent observed in pattern A solutions.

Network Diagram Construction

For all three types of tasks (design, implementation and troubleshooting), students were asked to sketch a diagram of the network described in the text of each scenario before beginning the actual task required for the scenario. While this diagram construction task was done in the context of a larger scenario, the critical characteristics of performance were similar across scenario types, and the results are grouped together here. Based on the design determination that the sequence of procedures in constructing the diagram is not important evidence for the inferences to be made from the diagram task only the final diagram was collected during the CTA and no data was collected on sequence of diagramming processes. Therefore, this discussion does not address the correctness of procedure and instead focuses on the outcome characteristics of the paper and pencil diagrams students constructed during the CTA.

Overall, the diagrams students constructed varied by level of detail and accuracy. Many students were able to represent the network at the correct level of abstraction for the purpose of the scenario, including essential devices, connections between devices, and appropriate labels. Other students, however, failed to represent important network connections and/or devices. In addition, some students represented unnecessary detail by focusing on

irrelevant physical characteristics of devices and the buildings which housed them. The patterns of behavior for the diagram activity are described in terms of a single observable feature: Correctness of Outcome. This Correctness of Outcome is composed of:

a) Extraneous Components

b) Necessary components

On this basis three patterns of performance were distinguished among the solutions to the network diagram problems in the CTA:

- **Network Diagram Pattern A.** This pattern of performance was characterized by the omission of essential features of the network.
- **Network Diagram Pattern B.** This pattern of performance was characterized by the inclusion of all essential characteristics without the addition of erroneous diagram elements. Furthermore, this pattern was also characterized by a restriction of the detail of the network representation to the necessary level of abstraction for the required tasks.
- **Network Diagram Pattern C.** This pattern of performance was characterized by the inclusion of all essential characteristics without the addition of erroneous diagram elements.

Network diagram pattern A: correctness of outcome.

Pattern A performance was characterized by the omission of key aspects of the network such as network devices, necessary connections between devices or addressing and other key diagram labels. The solution may, for example, fail to connect two routers that must have a physical connection between them. Diagrams may also show additional incorrect elements such at the wrong devices (e.g. a hub instead of a switch) or include incorrect connections or labels (e.g. wrong IP address). Finally, the diagram may also contain a number of extraneous or irrelevant details which could not be an influential factor in the required scenario task.

Network diagram pattern B: correctness of outcome.

Pattern B solutions were characterized by the provision of all or nearly all of the key features of the network (e.g. devices, connections, address labels) in an accurate manner, without the inclusion of erroneous elements or relationships among network components. Furthermore, the network was represented at the correct level of abstraction and the representation contained little, if any, irrelevant detail.

Network diagram pattern C: correctness of outcome.

Pattern C solutions provided all or nearly all of the key features of the network (e.g. devices, connections, address labels) in an accurate manner, without the inclusion of erroneous elements or relationships among network components. However, despite the accuracy of the network representation, this pattern of solution also has a tendency to include a substantial number of irrelevant details about the network that are not germane to the networking task.

These patterns of performance for the network diagram task may be summarized by saying that Pattern A diagrams are ineffective; Pattern B diagrams are both effective and efficient, that is, efficacious—which is a common feature of expert representations (Kindfield, 1999);—and Pattern C diagrams are effective but not efficient.

Summary

The results revealed a degree of similarity in the overall classification structure across tasks. For example, the Correctness of Outcome is a universal characteristic of evaluation across the task types and Correctness of Procedure is common, though not universal. However, while the general classifications (e.g. Correctness of Outcome) can be consistent across task types, the characteristics of solutions that contribute to those classifications are specific to the tasks, as is evident in the more fine-grained characteristics for each task type. The observed patterns of performance for all task types were consistent with the three primary classifications of performance patterns, with one (pattern A) indicative of an inadequate pattern of performance, one (pattern B) indicative of a high level of performance, and one (pattern C) indicative of a moderately successful performance. These patterns of performance may be generally summarized by saying that Pattern A is ineffective; Pattern B is both effective and efficient, that is, efficacious—a common feature of expert representations (Kindfield, 1999)– and Pattern C is somewhat effective but not efficient.

DOMAIN MODEL

The CTA results, in combination with other analyses of the domain (from source materials, documentation, interviews with SMEs, etc.), guide the design and implementation of the Student Model, Evidence Models, and Task Models for the networking prototype assessment. Again we emphasize that these models do not simply follow from domain analysis results, but rather from a process begun in advance of the CTA and iteratively revisited and revised from initial domain investigations through the CTA—and which, in fact, continues into the empirical estimation of assessment characteristics. While a full explication of model development is beyond the scope of this paper, the

following sections discuss a sampling of the implications of the CTA, in combination with the preceding analyses of the domain, for the construction of these ECD models. A full treatment of this model development process, including the statistical model, will be addressed in future work.

Implications for the Development of a Student Model

The domain analysis, including the CTA, culminated in a Domain Model (sometimes called an unrestricted student model) representing the constellation of proficiencies and related claims important for success as a CNAP student. The resultant Domain Model is provided as Figure 2 (as are a few associated claims, indicated by stars), representing not only the knowledge, skill and ability necessary, but also the dependencies among them. For example, a student's Network Proficiency (which consists of five interrelated skills) is modeled to be dependent on their Domain Disciplinary Knowledge (which also consists of multiple components). In addition, Design and Troubleshooting are modeled to be conditionally independent given Network Proficiency. Further explication of Design is documented through its associated claims(s), two of which (128a and 400) are provided as examples. In the latter stages of assessment design, the development of reporting rules for a student model further specify, through functions of the posterior distributions for student-model variables, how the values of one or more of these variables relate to particular claims. This Domain Model will later comprise the foundation for the Student Model for the NetPass assessment.

The Domain Model is composed of a number of variables representing aspects of knowledge, skill and ability. The Domain Disciplinary Knowledge variable represents the declarative knowledge of network components and operation, and therefore is the type of knowledge typically assessed through tasks requiring the recall of specific elements of network knowledge (e.g. multiple-choice). There are a number of elements of declarative knowledge represented as part of the Domain Disciplinary Knowledge, including: the OSI network model; addressing schemes; hardware components of a network; media; IOS; protocols; and security.

The Network Proficiency variable represents the skills and procedural knowledge (as opposed to declarative knowledge) necessary to perform a variety of tasks critical to successful network operations. These network operations include several skills: Implementing (configuring) a network; Troubleshooting; Designing; Operating; and Planning. (Only Troubleshooting, Implementation and Design tasks were utilized in the CTA and the prototype assessment). As each of these network activities requires declarative knowledge in order to conduct the procedures

required to perform these tasks, there is a modeled relationship between the declarative knowledge represented in Domain Disciplinary Knowledge and the procedural knowledge required for Network Proficiency.

The Network Modeling variable is the ability of the student to represent a network structure that may facilitate their Network Proficiency in various types of tasks. The ability to produce a model of a network requires, in part, Domain Disciplinary Knowledge, which is therefore represented as a prerequisite of Network Modeling ability. Since the ability to produce a model of the network in question is thought to facilitate such tasks as troubleshooting and design, there is a modeled relationship between this ability and the Network Proficiency ability.

Implications for Development of Evidence Models

By utilizing tasks in the CTA that were explicitly designed to yield evidence relevant to anticipated assessment claims the CTA results yielded specific examples of evidence supporting these claims. From the CTA, we have observable examples of errors and misconceptions of students at various levels of expertise, as well as the elements they successfully implement. This evidence will inform development of the Evidence Models, which relate students' actions to the states of the Student Model variables. The Evidence Model is the vehicle by which the raw data from the student's solution is transformed into evidence about the specific claims to be made about the student, and absorbed into the Student Model to update our understanding of the student knowledge, skills and abilities. These Evidence Models consist of two subcomponents: the evidence rules and the statistical model. The evidence rules provide for the representation of work product elements as variables (called *observables*) that are used as evidence in the evidence accumulation process. Their values are summaries of the lower level features of the assessment data compiled in a manner similar to Clauser et al. (1995). The statistical model for the assessment takes the values of these fine-grained observables from student performances and uses them to update the values of student model variables. The structure of the Domain Model has implications for the required characteristics and capabilities of the statistical model used to link the values of observables to the estimates of Student Model variables.

Implications for Task Models

The results of the CTA also serve to suggest required task characteristics to maximize their tendency to elicit evidence distinguishing among students, thus fulfilling the summary reporting and diagnostic feedback requirements and ensuring that the evidential requirements for the assessment claims can be fully met. As the CTA provides an opportunity to observe actual student behavior and strategies as they work through the complex tasks this

provides an ideal opportunity to determine which aspects of the tasks are serving well and which task elements might be altered to yield better evidence of student ability, thereby better supporting the inferences made about student model variables. As such, the findings of the CTA have several implications for task design in honing the tasks into a better tool for updating the student model variables in the prototype implementation.

One aspect of the Domain Model with implications for the tasks used in assessment is Domain Disciplinary Knowledge. This proficiency correlates with the Network Proficiency student model variable. As our estimate of students' Network Proficiency increases, so does our estimate of their Domain Disciplinary Knowledge. There is, however, a potential ambiguity in the model. If students encounter difficulties with tasks we cannot be certain, based on the network simulations alone, whether those difficulties are due to deficiencies in Domain Disciplinary Knowledge or a lack of Network Proficiency. To disambiguate estimates of these proficiencies direct evidence of Domain Disciplinary Knowledge could be obtained through the inclusion of items designed to specifically target this declarative knowledge in addition to the skills and abilities targeted by the design of simulation tasks.

Similarly, Error Identification (part of the Correctness of Outcome from the troubleshooting tasks in the CTA, see above) in troubleshooting tasks is based on the observation that the student successfully remedies an existing network error. By observing that the student remedies an existing error this observation is used as evidence that the student has both identified the presence of the error and understands the procedures for remedying the network error. However, the possibility exists that a student has sufficient skills to identify an error in a network but insufficient skill to actually intervene and correct the deficiency. The initial evidence identification process did not attempt to disambiguate these two possible factors (error identification and error correction) contributing to the Error Identification variable. As a result there would be implications of this potential ambiguity for assessment design, possibly requiring the addition of a task allowing for this ability to disambiguate deficiencies in ability.

CONCLUSION

The principles of ECD have provided a framework for conducting critical tasks in assessment design, beginning with an analysis of the domain and an integral CTA, that has resulted in an appropriate complex collection of design characteristics to guide the construction and implementation of a complex assessment design. This approach, while advantageous for any assessment design, has been instrumental in permitting the construction of a complex measurement model that maintains a clear and fundamental relationship between the assessment goals and

the critical elements of the assessment, from the initial analysis of the domain through the identification of characteristics that will guide the development of the student, evidence and task models and the implementation of the four process assessment architecture. Specifically, this approach has set the goals for the assessment, established a process for analysis of the domain, provided for explicit claims about the examinee that the assessment must support, elaborated the manner in which the observable variables from task performance provide evidence to support claims about the examinee, directed the conduct of a CTA to determine the appropriate design of the assessment models, and provided for the construction of the necessary models on the basis of evidence obtained in the investigations.

Now that the necessary work for understanding the assessment design has been completed, the next phase of the project will describe the models driving assessment production and operational implementation. This phase addresses the specification of student models, evidence models and task models that comprise the assessment design, as well as initial statistical specification for the Bayesian network serving as the NetPass statistical model. The next phase will include field trials with CNA students throughout the CLI educational network. From these field trials it is anticipated that opportunities for improving the interface, task presentation, and other aspects of the assessment will become evident. These field trials also provide initial evidence to accumulate the empirical basis for the conditional probabilities used in the Bayesian networks comprising the statistical model for the assessment. These will be further refined by estimation procedures developed from techniques described in Mislevy, Almond, Yan, and Steinberg (1999). The implications of this phase of development will be elaborated with respect to the four-process assessment delivery architecture

"We live in an age when we are still more adept at gathering, transmitting, storing, and retrieving information than we are at putting this information to use in drawing conclusions from it," assert statistician Jay Kadane and evidence scholar David Schum (1996, p. xiv). This is surely the case with the complex assessments now appearing in operational practice—whether in portfolios, extended projects, or computer-based simulations—all offering the promise of evidence about skills and knowledge beyond that captured in traditional assessments, yet each lying tantalizingly beyond the firm grasp of traditional assessment practices.

Tasks in standardized tests are encapsulated and observations are spare mainly because, historically, our delivery processes in our assessment designs could not accommodate more complex forms of evidence. The evolution of computer technology and simulation capabilities shatter this barrier—only to reveal a new one: just how to shape our design of the assessments to make sense of the "rich and realistic data" now obtainable. A principled

approach to this problem requires insights and methods from several disparate fields, including technology, cognitive psychology, educational measurement, and the subject domain in question—a daunting challenge resulting in a complex model for dealing with the rich evidence provided by such data. This paper has provided an example of how ECD provides just such an approach, developed in a replicable way and using a meaningful example in computer networking, to create a complex measurement model capable of supporting the use rich and realistic student performances.

References

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2001). A sample assessment using the four process framework. *CSE Technical Report #543*. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Bejar, I.I. (2002). Generative testing: From conception to implementation. In S.H. Irvine & P.C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Hillsdale, NJ: Erlbaum.

Bennett, R. E. & Bejar I. I. (1998). Validity and automated scoring: It's not only the scoring. Educational Measurement: Issues and Practice, 17(4), 9-17.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new structure of intellect" In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27-57). Hillsdale, NJ: Erlbaum.

Clauser, B.E., Subhiyah, R., Nungester, R.J., Ripkey, D., Clyman, S.G., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. *Journal of Educational Measurement, 32*, 397-415.

Ericsson, K. A., & Smith, J., (1991). Prospects and limits of the empirical study of expertise: An introduction. In K.A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits*. Cambridge: Cambridge University Press.

Gitomer, D.H., Steinberg, L.S., & Mislevy, R.J. (1995). Diagnostic assessment of trouble-shooting skill in an intelligent tutoring system. In P. Nichols, S. Chipman, & R. Brennan (Eds.), Cognitively diagnostic assessment (pp. 73-101). Hillsdale, NJ: Erlbaum.

Irvine, S.H., & Kyllonen, P.C. (Eds.) (2002). *Item generation for test development*. Hillsdale, NJ: Erlbaum.

Jensen, F.V. (1996). An Introduction to Bayesian Networks. New York, NY: Springer-Verlag New York, Inc.

Kadane, J.B., & Schum, D.A. (1996). *A probabilistic analysis of the Sacco and Vanzetti evidence*. New York: Wiley.

Kindfield, A.C.H. (1999). Generating and using diagrams to learn and reason about biological processes. *Journal of the Structure of Learning and Intelligent Systems, 14(2),* 81-124.

Klein, M.F., Birenbaum, M., Standiford, S.N., & Tatsuoka, K.K. (1981). *Logical error analysis and construction of tests to diagnose student "bugs" in addition and subtraction of fractions*. Research Report 81-6. Urbana, IL: Computer-based Education Research Laboratory, University of Illinois.

Melnick, D. (1996). The experience of the National Board of Medical Examiners. In E.L. Mancall, P.G. Vashook, & J.L. Dockery (Eds.), *Computer-based examinations for board certification* (pp. 111-120). Evanston, IL: American Board of Medical Specialties.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R.J., (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R.J., Almond, R.G., Yan, D., & Steinberg, L.S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (437-446). San Francisco: Morgan Kaufmann Publishers, Inc.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (pp. 97-128). Hillsdale, NJ: Erlbaum.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (in press). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Discussion.*

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior, 15,* 335-374.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (in press). Making sense of data from complex assessments. *Applied Measurement in Education.*

Newell, A., & Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Steinberg, L.S., & Gitomer, D.H., (1993). Cognitive task analysis and interface design in a technical troubleshooting domain. *Knowledge-Based Systems,6,* 249-257.

Steinberg, L.S., & Gitomer, D.G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24,* 223-258.

Steinberg, L., Mislvey,R., Almond, R., Baird, A., Cahallan, C., Chernick, H., Dibello, L., Kindfield, A., Senturk, D., Yan, Duanli (2000). Using evidence-centered design methodology to design a standards-based learning assessment. Research Report. Educational Testing Service. Princeton, NJ.

Whitely, S.E. [Embretson, S.E.] (1976). Solving verbal analogies: Some cognitive components of intelligence test items. Journal of Educational Psychology, 68, 234-242.

Acknowledgements

This paper is based on research conducted jointly by the Cisco Learning Institute (CLI) and Educational Testing Service (ETS). The principal investigators of the project are John Behrens, Cisco Systems, and Malcolm Bauer, ETS. Additional members of the research team whose contributions we gratefully acknowledge include, from Cisco, Tara Jennings, Sarah Demark, Ken Stanley, Bill Frey, Michael Faron; from ETS, Peggy Redman; and from Unicon, Perry Reinhert. We would like to express our gratitude to the 27 students who volunteered to work through the cases of the cognitive task analysis, and their instructors, Mike Hogan, Dara West, K Kirkendall, and Tenette Petelinkar. We especially wish to acknowledge and thank the Cisco Learning Institute for funding this research.

Table 1: Example Claim and Evidence

Claim: Identify the cause of connectivity problems at the physical, data link, and network layers of the OSI model

Representations to capture information from student:

1. Log file of IOS commands
2. Configuration files for routers (state of network)
3. Worksheet (set of faults)
4. Network Diagram
5. Essay

Table 2

Observable Features (Evidence) from Specific Representations

Observable Feature	*General category*	Representations
1. Steps taken to identify problem(s)	*Correctness of procedure*	1
2. Identification of network problem(s)	*Correctness of outcome*	*1,2,3*
3. Connection between steps and problem(s)	*Connections between procedure and outcome*	*1,3,5*
4. Problem solving logic used for determining cause	*Rationale for procedure*	*5*

Figure Captions

Figure 1. The three central models of an Evidence-Centered Design framework.

Figure 2. Networking Proficiencies and Claims

Figure 1

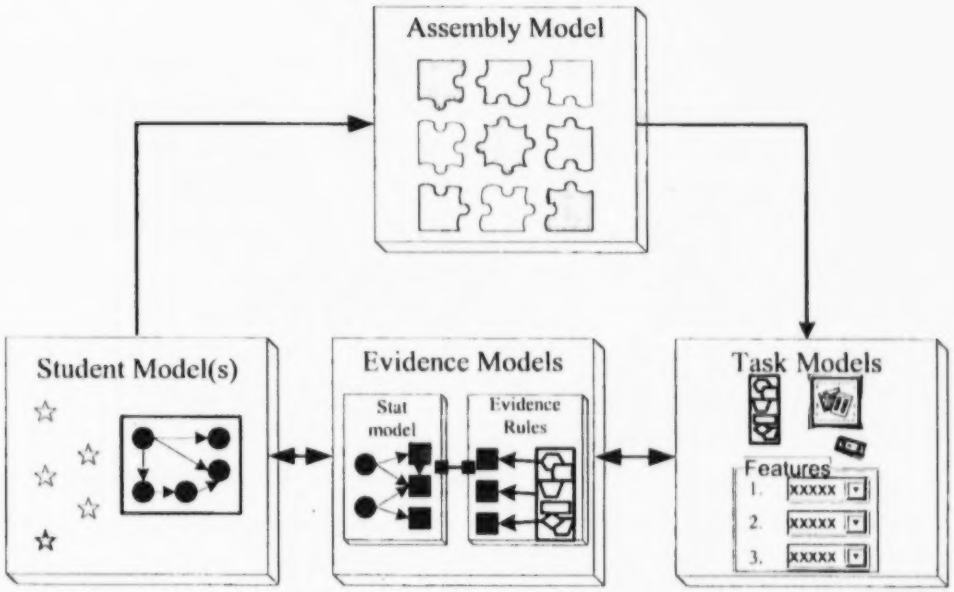
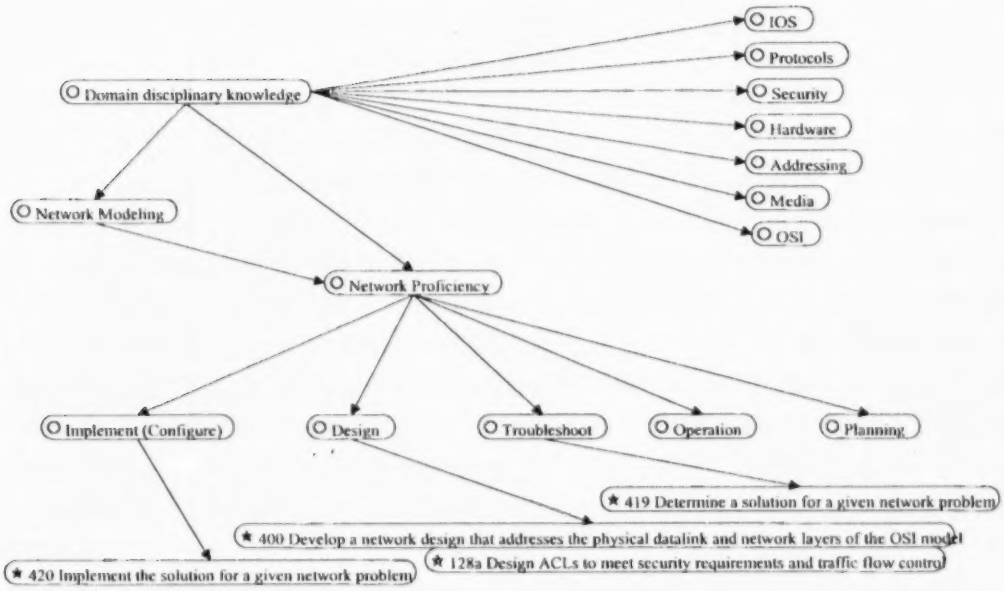

Assembly Model
Student Model(s)
Evidence Models
Stat model
Evidence Rules
Task Models
Features
1. xxxxx
2. xxxxx
3. xxxxx

Figure 2. Networking Proficiencies and Claims



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM035023

I. DOCUMENT IDENTIFICATION:

Title: Creating a Complex Measurement Model Using Evidence Centered Design

Author(s): David M. Williamson; Malcolm Bauer; Linda Steinberg; Robert J. Mislevy; John T. Behrens

Corporate Source:	Publication Date:
ETS	April 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in th monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, a electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproductio release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the botto of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: David M. Williamson	Printed Name/Position/Title: David M. Williamson, Research Scientist	
Organization/Address: ETS, Rosedale Road, Princeton, NJ 08541	Telephone: 609-734-1303	FAX:
	E-Mail Address: dmwilliamson@ets.org	Date: 6/4/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse: **University of Maryland**
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions